E-MAIL FILTERING USING BAYESIAN NETWORK

A thesis submitted to the Graduate School in partial
fulfillment of the requirement for the degree
Master of Science Intelligence System (IS)
Universiti Utara Malaysia

By
Kanakorn Horsiritham

JABATAN HAL EHWAL AKADEMIK
(*Department of Academic Affairs*)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

KANAKORN HORSIRITHAM

calon untuk Ijazah
(candidate for the degree of) **MSc. (Int. Sys)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project paper of the following title)

E-MAIL FILTERING USING BAYESIAN NETWORK

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the filed is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **MISS NOORAINI YUSOFF**

Tandatangan
(Signature) : Hooraini

Tarikh
(Date) : 17 OCTOBER 2004

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of the Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean of Graduate School

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman.

ABSTRACT

E-Mail is important today. It is applied in many application; Education, Business and personal communication. Once there are too many E-Mail arrived in the mailbox and mostly are unwanted E-Mail, called Spam. Spam is a costly problem. At Prince of Songkhla University (PSU), there are around 5,000 e-mail users and around 40,000 messages received a day. There are 10 % of them are virus and spam messages. Otherwise, the mail server has to pay memory and CPU load to process these virus and spam messages. These may cause the server response slowly and sometime once the system resources are insufficient, the mail server may crash and unavailable. Many filtering techniques are proposed. Bayesian Network is one of the popular Spam Filtering methods. This project is study Bayesian Network using SpamBayes, Open Source Software. Spam E-Mail are always written in English but at PSU there are Thai Language Spam found increasingly. Thai Language is different from English Language because English word is separated by space but Thai Language is not. The project examines the SpamBayes accuracy on Spam classification of mix Thai and English E-Mail messages. Thai and English E-Mail are trained together and test messages are also Thai and English mixed. The result shows that SpamBayes can classify Spam both in Thai or English.

ACKNOWLEDGMENTS

I would like to thank Professor Fadzilah Siraj for his guidance and patience in the past year in the completion of this project. And also Nooraini and Azizi too.

Thanks Mr. Wipat Srutiprom, Computer Center, Prince of Songkhla University, Thailand for knowledge and every techniques. Thanks Dr. Nittida Nualsri for knowledge about UNIX and for her idea on this project. Thanks Dr. Prawat Wetprasit and Dr. Wipada Wetprasit for guidance in English.

I would also like to thank Sasalak and Aumnat Tongkaw who were both my sister, brother and the best consultant for everything.

Thank you, Mam, Saranee who encourage me every time I were tired and nearly give up. Thanks for my family, Mom and Grand Mom who take care about my health. Thanks to Jedt, my geek friend too.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

CHAPTER 1

INTRODUCTION

This chapter presents the overview of the E-Mail Filtering using Bayesian Network. The project objective, significant and scope of study are also discussed.

1.1 Overview

E-Mail is important today. From the day when computers were first linked together through some form of a network, computer users have been sending messages to each other over the wires. Now with the worldwide presence of the Internet, computer networks handle trillions of messages every day. Electronic mail, or e-mail, is one of the most commonly used services on computer networks and the Internet. The major attraction of e-mail is its almost immediate delivery. Despite the distance between the sender and the receiver, an e-mail message can find its way anywhere in the world within minutes. E-mail has a way of drawing the global community closer together.

E-mail More Important Than the Phone In Business. A survey of businesspeople at 387 organizations found that 80 percent believed e-mail was more important than the telephone in communicating with coworkers, customers, or partners. Just as

The contents of the thesis is for internal user only

REFERENCE

Androutsopoulos I., Koutsias J., Chandrinos K. & Spyropoulos C. D. (2000). *An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages*, SIGIR 160-167.

Anthony B. (2004). *SpamBayes–Background Reading*, <http://spambayes.sourceforge.net/background.html> (20 September 2004)

Bauer M. & Winter B. (2000). *Using Postfix for Secure SMTP Gateways*, Linux Journal, volume 2000, Issue 78es.

Bevilacqua-Linn M. (2003). *Machine Learning for Naïve Bayesian Spam Filter Tokenization*, University of Rochester, New York.

Bickmore, T. W. (1994). *Real-Time Sensor Data Validation*, NASA Contractor Report 195295, National Aeronatics and Space Administration.

Cranor L. F. & LaMacchia B.A. (1998). *Spam!* Communications of the ACM, Vol. 41, No. 8, p.74 – 83.

Cunningham P., Nowlan N., Delany SJ. & Haahr M. (2003). *A Case-Based Approach to Spam Filtering that Can Track Concept Drift*, The ICCBR'03 Workshop on Long-Lived CBR Systems, Trondheim, Norway.

Diao Y., Lu H. & Wu D. (2000). *A Comparative Study of Classification Based Personal E-mail Filtering*, Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, p. 408 – 419. Springer-Verlag London, UK.

Elkan C. (1997). *Naïve Bayesian Learning*, Department of Computer Science, Harvard University.

Graham P. (2002). *A Plan for Spam,* <http://www.paulgraham.com/spam.html> (20 September 2004)

Huang, T., Koller, D., Malik, J., Ogasawara, G., Rao, B., Russell, S., & Weber, J. (1994). *Automatic Symbolic Traffic Scene Analysis Using Belief Networks,* Proceedings of National Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA.

Hung E. (2001). *Deduction of Procmail Recipes from Classified Emails*, Department of Computer Science, University of Maryland.

Itskevitch J. (2001). *Automatic Hierarchical E-Mail Classification Using Association Rules,* Simon Fraser University.

Kevin R. G. (2003). *Using Latent Semantic Indexing to Filter Spam,* SAC, ACM.

Kristian E. (2004). *Winning the War on spam: Comparison of Bayesian spam filters,* <http://home.dataparty.no/kristian/reviews/bayesian/> (20 September 2004)

Meyer T.A.& Whateley B. (2004). *SpamBayes: Effective open-source, Bayesian based, email classification system,* CEAS, Canada.

Niedermayer D. (1998). *An Introduction to Bayesian Networks and their Contemporary Applications*
, <http://www.niedermayer.ca/papers/bayesian/index.html> (20 September 2004)

Pazzani J. M. (2000). *Representation of Electronic Mail Filtering Profiles: A User Study*, Department of Information and Computer Science, University of California.

Peter T. (2004). *SpamBayes–Credit*, <http://spambayes.sourceforge.net > (20 September 2004)

Redmond M. and Adelson B. (1998). *AlterEgo E-Mail Filtering Agent - Using CBR as a Service*, In "Case-Based Reasoning Integrations, Papers from the 1998 Workshop" (AAAI-98). 143-148. Madison, WI. AAAI Press.

Rennie J. D. M. (2000). ifile: *An Application of Machine Learning to Email Filtering,* AI Lab, MIT, KDD2000 Text Mining Workshop Boston, MA USA.

Robinson G. (2003). *Better Bayesian Filtering* , <http://www.paulgraham.com/better.html> (20 September 2004)

Sabil M. (2002). *MeatSlicer: Spam Classiffcation with Naieve Bayes and Smart Heuristics*, <http://web.mit.edu/msalib/www/writings/classes/6.034/project2/paper.pdf> (20 September 2004)

Sahami M., Dumais S., Heckerman D., and Horvitz E. (1998) *A Bayesian approach to filtering junk email*, In Proceedings of the AAAI Workshop on Learning for Text Categorization.

Vemuri V. and Tang N. (2004). *Solving Inverse Problems via Machine Learning and Knowledge Discovery*, In (Eds. Takumi Ichimura and Katsumi Yoshida.), Knowledge-Based Intelligent Systems for Healthcare, CRC Press.

White Paper. (2003). *"Symantec: Neural Network-based Antispam heuristics"* <http://www.symantec.com> . (16 July 2004)